

# Computation and Visualisation for large-scale Gaussian updates

Jonathan Rougier\*  
School of Mathematics  
University of Bristol

Andrew Zammit Mangion and Nana Schoen  
School of Geographical Sciences  
University of Bristol

## Abstract

In geostatistics, and also in other applications in science and engineering, we are now performing updates on Gaussian process models with many thousands or even millions of components. These large-scale inferences involve computational challenges, because the updating equations cannot be solved as written, owing to the size and cost of the matrix operations. They also involve representational challenges, to account for judgements of heterogeneity concerning the underlying fields, and diverse sources of observations.

Diagnostics are particularly valuable in this situation. We present a diagnostic and visualisation tool for large-scale Gaussian updates, the ‘medal plot’. This shows the updated uncertainty for each observation, and also summarises the sharing of information across observations, as a proxy for the sharing of information across the state vector. It allows us to ‘sanity-check’ the code implementing the update, but it can also reveal unexpected features in our modelling. We discuss computational issues for large-scale updates, and we illustrate with an application to assess mass trends in the Antarctic Ice Sheet.

**KEYWORDS:** VARIANCE UPDATE, VARIANCE BOUND, MEDAL PLOT, PRINCIPLE OF STABLE INFERENCE

---

\*Reader in Statistics. Address: School of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK. Email [j.c.rougier@bristol.ac.uk](mailto:j.c.rougier@bristol.ac.uk).

# 1 Introduction

Statisticians are now attempting inferences of a scale and complexity that were unthinkable even a few years ago. This is for a number of reasons:

1. Computers are more powerful, and have larger memories,
2. New statistical techniques are available to represent judgements on large collections of random quantities, and to compute on those judgements,
3. Large new datasets, including from remote sensing, are becoming available, and
4. There is a political need, and research funding, to address inference for complex systems, notably environmental systems.

Similar assessments have been given by Kalnay (2002, ch. 1, concerning meteorology) and Smith (2010, ch. 1, concerning decision support). In our application, outlined below, the state vector has about  $10^5$ , and there are  $3.5 \times 10^5$  observations. Statistical inferences of this scale are most easily handled using a Gaussian process prior, and the linearisation of the observation operator; or else the use of an optimisation approach that comes to very much the same thing (e.g., as in data assimilation for meteorology, see Apte *et al.*, 2008).

One concern in a complex inference is to verify that the code and the model are performing sensibly, and a second is to visualise the assimilation of very large numbers of observations. The latter is particularly challenging over multiple interacting processes. We present a visual diagnostic which meets both of these needs, based on an upper bound on the updated variance. It is almost obvious that the updated variance of any measured linear combination of the state vector has to be no larger than the smaller of its initial variance and the observation error variance. A scalar version of this result was the basis for L.J. Savage's principle of stable inference (Savage *et al.*, 1962; Edwards *et al.*, 1963), in which an observation of  $Y$  with error variance  $\tau^2$  had the effect of setting the updated variance of  $Y$  to  $\tau^2$ , regardless of the initial variance, provided only that the initial distribution was suitably flat in the region around the observation. Savage used this as a demonstration of how subjective Bayesian scientists might end up agreeing despite their different initial judgements. Below, we prove a similar result for a collection of observations (section 2.3).

The simplest use of our result is to trap errors, since violation of the upper bound is *prime facie* evidence of a failure in the code. Such failures are

unlikely to occur for moderately-sized problems, for which the matrix algebra can be programmed more-or-less as written, but are an issue for large problems, where the calculation must be broken into chunks, or approximated, e.g. using iterative solvers. But there is additional information available in the source of the bound (initial variance or observation error variance), and in the relation of the updated variance to its bound. This leads naturally to a visualisation tool in updates of random fields, for which the linear combinations are often localised in the domain. As no reference is made to the value of the observations, this diagnostic can be used before the observations are made available, for example in experimental design (e.g. Krause *et al.*, 2008).

Section 2 describes the theoretical result and its implications, its implementation, and the ‘medal plot’ for visualisation. Section 3 discusses computation for large-scale applications, including a powerful result concerning sparsity. Section 4 illustrates with an inference for mass trends in the Antarctic Ice Sheet.

## 2 Result and implications

### 2.1 An upper bound for the updated variance

Let  $\mathbf{X}$  be the collection of Gaussian random quantities, and  $\mathbf{Y} := A\mathbf{X}$  be the known linear combinations which are measured, where  $A$  is sometimes termed the ‘incidence matrix’. Let  $\mathbf{Z} := \mathbf{Y} + \mathbf{E}$  be the observations, including observation error  $\mathbf{E}$ . Denote the variance matrix of  $\mathbf{Y}$  as  $\Sigma$ , and take the observation error to be independent of  $\mathbf{X}$ , with variance matrix  $T$ . If  $V$  and  $W$  are two variance matrices, then write  $V \leq W$  exactly when  $W - V$  is non-negative definite. Then we have the following result, which applies not just in the Gaussian case, but also for more general second-order updating, such as the Bayes linear approach described in Goldstein and Wooff (2007).

**Theorem 1.** *Let  $\Sigma^* := \text{Var}(\mathbf{Y} \mid \mathbf{Z})$  be the updated variance matrix of  $\mathbf{Y}$ . If  $\Sigma + T$  is non-singular, then  $\Sigma^* \leq \Sigma$  and  $\Sigma^* \leq T$ .*

One immediate use for this result is in code verification, but it is also the basis of further results below, which are used for visualisation.

*Proof.* As  $\text{Cov}(\mathbf{Y}, \mathbf{Z}) = \Sigma$  and  $\text{Var}(\mathbf{Z}) = \Sigma + T$ , the updated variance of  $\mathbf{Y}$  satisfies

$$\Sigma^* = \Sigma - \Sigma(\Sigma + T)^{-1}\Sigma \tag{1}$$

(see, e.g., Mardia *et al.*, 1979, chapter 3). Hence  $\Sigma^* \leq \Sigma$  because the second term on the righthand side of (1) is non-negative definite. If we can show that

$$\Sigma - \Sigma(\Sigma + T)^{-1}\Sigma = T - T(\Sigma + T)^{-1}T, \quad (2)$$

then (1) and the same reasoning implies that  $\Sigma^* \leq T$ , completing the proof. Start with the identities

$$\mathbf{0} = \begin{cases} \Sigma - \Sigma(\Sigma + T)^{-1}(\Sigma + T), \\ T - T(\Sigma + T)^{-1}(\Sigma + T). \end{cases} \quad (3)$$

Equating the two terms on the righthand side and rearranging gives

$$\Sigma - \Sigma(\Sigma + T)^{-1}\Sigma - \Sigma(\Sigma + T)^{-1}T = T - T(\Sigma + T)^{-1}T - T(\Sigma + T)^{-1}\Sigma.$$

But the final terms on each side of this expression are equal, because they are symmetric, and (2) is proved. [See Piziak and Odell (2007, section 1.2) for more general results of this type.]  $\square$

It is important that this result holds for singular  $\Sigma$ , provided that  $\Sigma + T$  is non-singular. This is because we may well have replications in the observations; e.g. the same component of  $\mathbf{X}$  observed several times. This would be represented as duplicate rows in  $A$ . Alongside replications, we may well have more observations than components of the state vector, e.g. if we have multiple instruments with overlapping footprints. This would be represented by an  $A$  with more rows than columns. In both of these cases

$$\Sigma = A \text{Var}(\mathbf{X})A^T$$

would be singular (non-negative definite but not positive definite). No matter what the form of  $A$ , a non-singular  $T$  is sufficient for  $\Sigma + T$  to be non-singular (positive definite). Thus Theorem 1 always holds if there is measurement error.

## 2.2 Local and global updating

A further useful result concerns the relationship between the *joint* update  $\text{Var}(Y_i | \mathbf{Z})$  and the *local* update  $\text{Var}(Y_i | Z_i)$ . This is a special case of the following more general result about nested updates.

**Theorem 2.** *Let  $B$  and  $B'$  be two subsets of the indices of  $\mathbf{Y}$ , with  $B \supset B'$ . Then*

$$\text{Var}(\mathbf{Y}_{B'} | \mathbf{Z}_B) \leq \text{Var}(\mathbf{Y}_{B'} | \mathbf{Z}_{B'}) \leq T_{B',B'}. \quad (4)$$

*Proof.* It is a standard result in second-order updating that  $B \supset B'$  implies that

$$\text{Var}(\mathbf{Y} \mid \mathbf{Z}_B) \leq \text{Var}(\mathbf{Y} \mid \mathbf{Z}_{B'}),$$

(see, e.g., Goldstein and Wooff, 2007, section 5.2). This implies the first inequality in (4), because  $\mathbf{Y}_{B'}$  is a set of linear combinations of  $\mathbf{Y}$ . Theorem 1 implies that  $\text{Var}(\mathbf{Y}_{B'} \mid \mathbf{Z}_{B'}) \leq T_{B',B'}$  for any  $B'$ , which is the second inequality in (4).  $\square$

The interesting feature of this result is the introduction of  $T$  in place of  $\Sigma$  in the second inequality, courtesy of Theorem 1. Often the measurement error variance  $T$  is much smaller than the initial variance  $\Sigma$ , and so Theorem 2 represents a substantial lowering of the upper bound on any updated variance.

The next result follows immediately from Theorem 2, taking  $B$  to be the complete set of indices and  $B' = \{i\}$ :

$$\text{Var}(Y_i \mid \mathbf{Z}) \leq \text{Var}(Y_i \mid Z_i) \leq T_{ii} \quad \text{for each } i. \quad (5)$$

This ordering of global, local, and observation error variances is used our visualisation tool, presented in section 2.4. We can verify the second inequality in (5) by direct calculation:

$$\text{Var}(Y_i \mid Z_i) = \Sigma_{ii} - \frac{\Sigma_{ii} \cdot \Sigma_{ii}}{\Sigma_{ii} + T_{ii}} = \frac{\Sigma_{ii} \cdot T_{ii}}{\Sigma_{ii} + T_{ii}} \leq T_{ii}. \quad (6)$$

This expression shows that there is a limit to how much relative effect a local update can have. Taking  $T_{ii} \leq \Sigma_{ii}$ , for concreteness,

$$\inf_{T_{ii} \leq \Sigma_{ii}} \frac{\text{Var}(Y_i \mid Z_i)}{T_{ii}} = \inf_{\kappa \leq 1} \frac{1}{1 + \kappa} = \frac{1}{1 + 1} = \frac{1}{2}. \quad (7)$$

In other words, information from  $Z_i$  alone can push the updated variance of  $Y_i$  down to half of its upper bound, and this occurs when  $\Sigma_{ii} = T_{ii}$ . Eq. (6) also shows that if  $T_{ii} \ll \Sigma_{ii}$  then  $\text{Var}(Y_i \mid Z_i) \approx T_{ii}$ . In other words, the variance of the local update tends to the observation error variance as the observation error variance becomes small relative to the initial variance.

## 2.3 Limiting behaviour

The case where one of  $\Sigma$  or  $T$  is much larger than the other occurs frequently in practice, and it is interesting to consider the limiting case where, for concreteness,  $T$  becomes vanishingly small relative to  $\Sigma$ . However, there is a difficulty with this case: if  $\Sigma$  is singular, then a ‘vanishingly small’  $T$  will

ultimately conflict with the requirement that  $\Sigma + T$  be non-singular. But, as explained in section 2.1, it is common for  $\Sigma$  to be singular. Therefore the following result has additional conditions relative to Theorems 1 and 2, but it is powerful when these conditions hold.

**Theorem 3.** *Let  $\Sigma$  and  $T$  both be non-singular, and define  $r := \|T\Sigma^{-1}\|$ , where  $\|\cdot\|$  is any induced  $p$ -norm. If  $r < 1$  then*

$$\|\Sigma^* - T\| \leq \frac{\|\Sigma^{-1}\| \|T\|^2}{1 - r}. \quad (8)$$

*Proof.* Start from (1) and the top branch of (3) to show that

$$\Sigma^* = \Sigma(\Sigma + T)^{-1}T.$$

Now under the conditions of the Theorem both  $\Sigma$  and  $T$  are non-singular, and this expression can be rearranged to show that

$$\Sigma^* = (\Sigma^{-1} + T^{-1})^{-1}$$

(see also Rue and Held, 2005, section 2.3.3). Then (8) follows from a standard result about inverses and perturbations (see, e.g., Golub and Van Loan, 1996, Theorem 2.3.4).  $\square$

In other words, if both  $\Sigma$  and  $T$  are non-singular then as  $T$  becomes small relative to  $\Sigma$ , so the updated variance converges to  $T$ . However, it is important to appreciate that  $T$  non-singular is not, on its own, sufficient for this convergence. This is seen in the following counter-example with a singular  $\Sigma$ :

$$\text{Var}(\mathbf{X}) = 10^6 \begin{pmatrix} 1.0 & 0.4 \\ 0.4 & 1.0 \end{pmatrix}, \quad A = \begin{pmatrix} 1.0 & 1.0 \\ 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}, \quad T = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.1 \end{pmatrix},$$

for which, informally,  $T \ll \Sigma = A \text{Var}(\mathbf{X}) A^T$ . But

$$\Sigma^* = \begin{pmatrix} 0.17 & 0.08 & 0.08 \\ 0.08 & 0.09 & -0.01 \\ 0.08 & -0.01 & 0.09 \end{pmatrix}.$$

It can be checked that  $\Sigma^* \leq T$ , as required by Theorem 1, but clearly  $\Sigma^* \not\approx T$ . This combination of a singular  $\Sigma$  with both ‘large footprint’ imprecise observations and ‘small footprint’ precise observations occurs in our illustration in section 4.

**Principle of stable inference.** Theorems 1 and 3 provide a multivariate generalisation of L.J. Savage’s principle of stable inference, mentioned in the Introduction. We restate it here. The updated variance in this statement is a general second-order update, but applies in particular when  $\mathbf{Y}$  and  $\mathbf{E}$  are both Gaussian.

**Theorem 4** (Principle of stable inference, multivariate). *Let  $\mathbf{Y}$  be a vector of random quantities and  $\mathbf{Z}$  be a vector of noisy observations on the components of  $\mathbf{Y}$ . Define the measurement error as  $\mathbf{E} := \mathbf{Z} - \mathbf{Y}$ . Let  $\mathbf{E} \perp\!\!\!\perp \mathbf{Y}$  and  $\text{Var}(\mathbf{E})$  be non-singular. Then  $\text{Var}(\mathbf{Y}|\mathbf{Z}) \leq \text{Var}(\mathbf{E})$ . Furthermore, if  $\text{Var}(\mathbf{Y})$  is non-singular and  $\text{Var}(\mathbf{E}) \ll \text{Var}(\mathbf{Y})$  then  $\text{Var}(\mathbf{Y} | \mathbf{Z}) \approx \text{Var}(\mathbf{E})$ .*

This principle underlies the common ‘plug-in’ approximation

$$\text{truth} | \text{measurement} = \text{measurement} + \text{measurement error}.$$

Our results indicate that the critical modelling judgement under which this approximation provides a conservative or approximate assessment of uncertainty about the true values  $\mathbf{Y}$  is that the (additive) measurement error  $\mathbf{E}$  is probabilistically independent of  $\mathbf{Y}$ . The other conditions seem much less demanding in practice.

## 2.4 Visualisation: the ‘medal plot’

We would like to visualise various features of the variance update, particularly for those components of  $\mathbf{Y}$  which correspond to spatial locations. These features include, for a specified  $Y_i$ : what the upper bound is, and where it comes from; what the updated variance is; and what contribution is made by observations other than  $Z_i$ .

These considerations suggest the following ‘medal plot’. Each  $Y_i$  is represented by a medal of three concentric disks of decreasing radius:

1. A red/blue disk representing the upper bound on the updated variance of  $Y_i$ , either *red* where the prior provides the upper bound, or *blue* where the observation error provides the upper bound.
2. A *white* disk representing the updated variance using  $Z_i$  alone (local update).
3. A *gold* disk representing the updated variance using all observations (joint update).

In all cases, the radius of the disk is proportional to the standard deviation.

The medals can be scaled so that when displayed on a map they do not overlap by more than is necessary to preserve the systematic spatial patterns. When there is an overlap, it is more effective to plot all of the red/blue disks first, and then to overplot with the white disks, and then with the gold disks. In some application, including our illustration below, it is more effective to use a semi-transparent light-blue than white, so that underlying map features are preserved.

We recapitulate the properties of each medal, based on the results of the previous subsections. First, the three disks must be nested, with gold inside white inside red/blue; see (5). Hence, each medal appears as a gold central disk, a white annulus, and a red/blue rim.

Second, the outer (red/blue) rim cannot be thicker than  $1 - 1/\sqrt{2} \approx 0.3$  of the total radius of the medal, because the update from the local observation alone cannot reduce the updated variance to less than one half of its upper bound; i.e. the white disk cannot have a radius less than  $1/\sqrt{2} \approx 0.7$  of the total radius; see (7). Moreover, a very thin rim indicates a large discrepancy between the initial variance and the measurement error variance. Thus a very thin blue rim indicates that the initial variance is far larger than the observation error variance.

Third, if both  $\Sigma$  and  $T$  are non-singular, then the combined thickness of the (red/blue) rim and the white annulus is compressed towards zero when one of the two initial variances ( $\Sigma$  or  $T$ ) dominates the other, because in this case the updated variance (gold disk) is almost the same as the upper bound (red/blue disk); see section 2.3.

For a given medal at location  $i$ , we might be particularly interested in the thickness of the white annulus. This thickness shows us how much of the update of  $Y_i$  is coming from observations other than  $Z_i$ , with a thick annulus showing that other observations are making a large contribution (i.e. driving the updated variance well below what is achieved by  $Z_i$  alone). When we compare the medals across the domain of the observations we can see at a glance how the spatial scale of the update varies, by comparing the widths of the white annuli.

## 2.5 Model parameters

In many cases, the second-order structure of  $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  is expressed conditionally on uncertain parameters  $\theta$ ; these might be spatial correlation lengths in  $\mathbf{X}$ , for example. But there is no data-free bound in these cases, because

$$\text{Var}(\mathbf{Y} \mid \mathbf{Z}) = \text{E}\{\text{Var}(\mathbf{Y} \mid \mathbf{Z}, \theta) \mid \mathbf{Z}\} + \text{Var}\{\text{E}(\mathbf{Y} \mid \mathbf{Z}, \theta) \mid \mathbf{Z}\},$$



and both terms will depend on the value of  $\mathbf{Z}$ , which affects the update of  $\theta$ , and also the value of  $E(\mathbf{Y} \mid \mathbf{Z}, \theta)$ . In this situation, where diagnostics and visualisation are the goal, it is necessary to plug-in a point value for  $\theta$ . Such a plug-in suffices to trap coding errors, because the lower bound must be respected for every value of  $\theta$ . If  $\theta$  is well-constrained by  $\mathbf{Z}$ , then the plug-in visualisation based on the posterior mean for  $\theta$  will only differ a little from that with  $\theta$  integrated out.

### 3 Computation

A full-scale update of the expectation and variance of a large-scale Gaussian random field with many observations can be prohibitively expensive. Here we provide details for the much cheaper option of only updating the diagonal components of the variance matrix of the observations, which is all that is required for the medal plot visualisation we described in section 2.4.

When the state vector  $\mathbf{X}$  is very large, standard calculations that start with the specification of  $\Xi := \text{Var}(\mathbf{X})$  are no longer feasible, because of the cost of holding the elements of  $\Xi$  in memory while computing products such as  $\text{Var}(\mathbf{Y}) = A\Xi A^T$ . In this case, one powerful strategy is to approximate the distribution of  $\mathbf{X}$  using a Gauss Markov random field (GMRF), as described in Rue and Held (2005) and Rue and Martino (2007). A GMRF representation for  $\mathbf{X}$  allows a very efficient calculation for the elements of a medal plot, as we now discuss.

In a GMRF, the variance parameter of  $\mathbf{X}$  is  $Q := \Xi^{-1}$ . A high degree of conditional independence (e.g. arising from judgements of smoothness, or from  $\mathbf{X}$  being the composition of several independent processes) implies that  $Q$  is sparse, with the location of its zeros being known. This in turn implies that the Choleski factor  $L$  is also sparse, where  $L$  is lower-triangular and  $LL^T = Q$ . There are efficient algorithms for computing the non-zero elements of  $L$  from  $Q$ ; see, for example, Rue and Held (2005, section 2.4).

The next stage towards finding  $\Xi$  would be to compute this from  $L$ : notionally  $\Xi = L^{-T}L^{-1}$ . The efficient algorithm for this stage is based on the Takahashi equation, formalised in the result of Erisman and Tinney (1975). If  $L_{ij} \neq 0$ , then  $\Xi_{ij}$  can be computed directly from  $L$  and values of  $\Xi_{pq}$  for which  $p \geq i$ ,  $q \geq j$ , and  $L_{pq} \neq 0$ . The  $(i, j)$  elements of  $\Xi$  for which  $L_{ij} \neq 0$  are termed the *sparse subset* of  $\Xi$ . The other elements of  $\Xi$  can also be computed, *but only if needed*. This final observation is crucial for efficiency in our application, because if  $A$  is sparse then many of the elements of  $\Xi$  that are not in the sparse subset are not needed for the medal plot.

For our medal plot we require the three vectors

$$\text{diag}(A\Xi A^T), \quad \text{diag}(A\Xi^* A^T), \quad \text{and} \quad \text{diag} T,$$

where  $\Xi^* := \text{Var}(\mathbf{X} \mid \mathbf{Z})$ . In fact, we may not want all components of these three vectors, if we only want to visualise a subset of the observations. In this case, we would simply drop rows from  $A$ , and the corresponding rows and columns from the observation error variance matrix  $T$ .

Consider the first vector, whose  $i$ th element is

$$[\text{diag}(A\Xi A^T)]_i = \sum_k \sum_j A_{ij} \cdot \Xi_{jk} \cdot A_{ik} = \sum_j \sum_k (A_{ij} \cdot A_{ik}) \cdot \Xi_{jk}.$$

Hence  $\Xi_{jk}$  influences the  $i$ th diagonal element when both  $A_{ij} \neq 0$  and  $A_{ik} \neq 0$ . Therefore  $\Xi_{jk}$  does not influence the vector of diagonal elements if, for every  $i$ , either  $A_{ij} = 0$  or  $A_{ik} = 0$ . Now suppose that all of the elements of  $A$  are non-negative. In this case

$$[A^T A]_{jk} = 0 \iff A_{ij} = 0 \text{ or } A_{ik} = 0 \text{ for every } i. \quad (9)$$

Therefore if  $[A^T A]_{jk} = 0$  then  $\Xi_{jk}$  is not required in order to compute the elements of  $\text{diag}(A\Xi A^T)$ . Hence if  $A$  is sparse, such that many elements of  $A^T A$  are zero, then only a few extra elements of  $\Xi$  will be needed, beyond the sparse subset. And if  $Q$  is sparse, then  $L$  is sparse and the sparse subset of  $\Xi$  is small.

Exactly the same sequence of operations applies for the updated variance  $\Xi^*$ , starting from the updated precision matrix

$$Q^* := Q + A^T T^{-1} A$$

(Rue and Held, 2005, section 2.3.3). This updated precision matrix is sparse if  $Q$  and  $A$  are sparse, and if  $T$  is diagonal, or block diagonal with a small bandwidth. But there is a much more powerful result if  $A$  is non-negative and  $T$  is diagonal, as we now show.

First, introduce a new operator,

$$\text{zeros}(A)_{ij} := \begin{cases} 0 & A_{ij} = 0 \\ 1 & \text{otherwise.} \end{cases}$$

Note that  $A$  non-negative implies that

$$[\text{zeros}(A^T A)]_{jk} = 0 \iff [A^T A]_{jk} = 0. \quad (10)$$

Now let  $T$  be diagonal, in which case

$$\text{zeros}(Q^*) = \text{zeros}(Q + A^T T^{-1} A) \stackrel{?}{\geq} \text{zeros}(A^T T^{-1} A) = \text{zeros}(A^T A).$$

The second inequality, marked as ‘ $\stackrel{?}{\geq}$ ’, needs a qualification. It is *not* the case that  $\text{zeros}(M + N) \geq \text{zeros}(N)$  for arbitrary  $M$  and  $N$ . However, as in Erisman and Tinney (1975), we will assume that every  $[Q + A^T T^{-1} A]_{ij}$  which must be computed is treated as nonzero, even if its value is zero due to numerical cancellation. The inequality follows if we adopt this treatment.

To continue, note that  $\text{zeros}(L^*) \geq \text{zeros}(Q^*)$  in the lower triangle of  $Q^*$ , where  $L^*$  is the lower triangular Choleski factor of  $Q^*$ ; see Rue and Held (2005, Corollary 2.2, section 2.4). Thus we have, for  $j \geq k$ ,

$$\text{zeros}(L^*)_{jk} \geq \text{zeros}(Q^*)_{jk} \geq \text{zeros}(A^T A)_{jk}.$$

Hence if  $\text{zeros}(L^*)_{jk} = 0$  then  $\text{zeros}(A^T A)_{jk} = 0$ , which implies that  $[A^T A]_{jk} = 0$  from (10), which implies that  $\Xi_{jk}^*$  is not required in order to compute the elements of  $\text{diag}(A \Xi^* A^T)$ , from (9). In summary, we have proved the following result.

**Theorem 5.** *If  $A$  is non-negative and  $T$  is diagonal, then the elements of  $\Xi^*$  required in order to compute  $\text{diag}(A \Xi^* A^T)$  are a subset of the sparse subset of  $\Xi^*$ .*

This result implies that the diagonal elements of  $A \Xi^* A^T$  are available ‘for free’ once we have computed the sparse subset of  $\Xi^*$ . The conditions on  $A$  and  $T$  are both very natural for large-scale spatial and spatial-temporal modelling. Indeed, they were a feature of our illustration well before we derived this result.

## 4 Illustration

Our illustration is part of a mass-balance estimate for the Antarctic Ice Sheet (AIS), which is the world’s largest freshwater reservoir. Here we provide the briefest outline of our inference, which we describe in detail elsewhere (Zammit-Mangion *et al.*, 2014a,b).

In order to determine the AIS contribution to sea-level change, we need to decompose the change in height of the AIS over a fixed time period into the sum of four main processes: change in the height of the underlying rock, effect of ice dynamics, firn compaction (densification of past years’ snow), and the net effect of surface processes (precipitation, run-off, melt, and refreeze).

Then to quantify the contribution to sea-level change, we sum the changes in height of ice, firn, and surface processes inside the grounding line (see the caption to Figure 1) over the AIS, and then map those to mass changes using specified densities.

We have observations from three types of instrument. First, a small number of GPS receivers on rocky outcrops, which give accurate observations for change in height of the underlying rock (at those outcrops). Second, satellite altimetry, which gives observations of height change (i.e. summing the four processes) along specified transects. Third, satellite gravimetry (Gravity Recovery and Climate Experiment, or GRACE), which provides measures of mass change, and therefore sees a linear combination of change in the height of the underlying rock, the ice dynamics and surface processes (firn compaction changes height but not mass). These three instruments have very different spatial footprints, with GPS being a point observation, altimetry having a footprint of  $\sim 1$  km (treated as a point observation), and gravimetry having a footprint of  $\sim 400$  km.

This is an inherently statistical problem because: (i) we have three instruments for four fields; (ii) there are substantial observation errors, (iii) the footprints of the instruments are of such different sizes, (iv) the observations do not cover the whole of the AIS, and (v) uncertainty assessment is a crucial output for impact studies related to sea-level rise. The problem becomes soluble once we incorporate prior information about the processes, notably their variabilities and their characteristic length scales, both of which can vary spatially. As well as the four fields, our unknowns include statistical parameters for the processes and in the observation equation.

For this illustration we used finite element basis functions to model each of the four processes (see, e.g., Lindgren *et al.*, 2011), with variable resolution to account for greater heterogeneity near the coastline. We used a blocked Gibbs sampler to update the processes conditional on the statistical parameters, and to update the statistical parameters conditional on the processes. Then we plugged in the maximum *a posteriori* estimate of the statistical parameters (which were well-constrained), and redid the update of the fields, to draw the medal plots. We illustrate with a medal plot for the gravimetry observations for 2006, shown in Figure 1. Recollect that the medals show the update from all observations; e.g. the gravimetry linear combinations are updated not just by the gravimetry observations, but also by GPS and altimetry.

We highlight some of the wealth of information in this figure. First, almost all of the medals have blue rims, showing that the upper bound on the updated variance comes from the observation error variance, not the initial variance. There is one exception, which is at about  $(+800 \text{ km}, +250 \text{ km})$ . Here our model for mass trends implies a small initial expectation and vari-

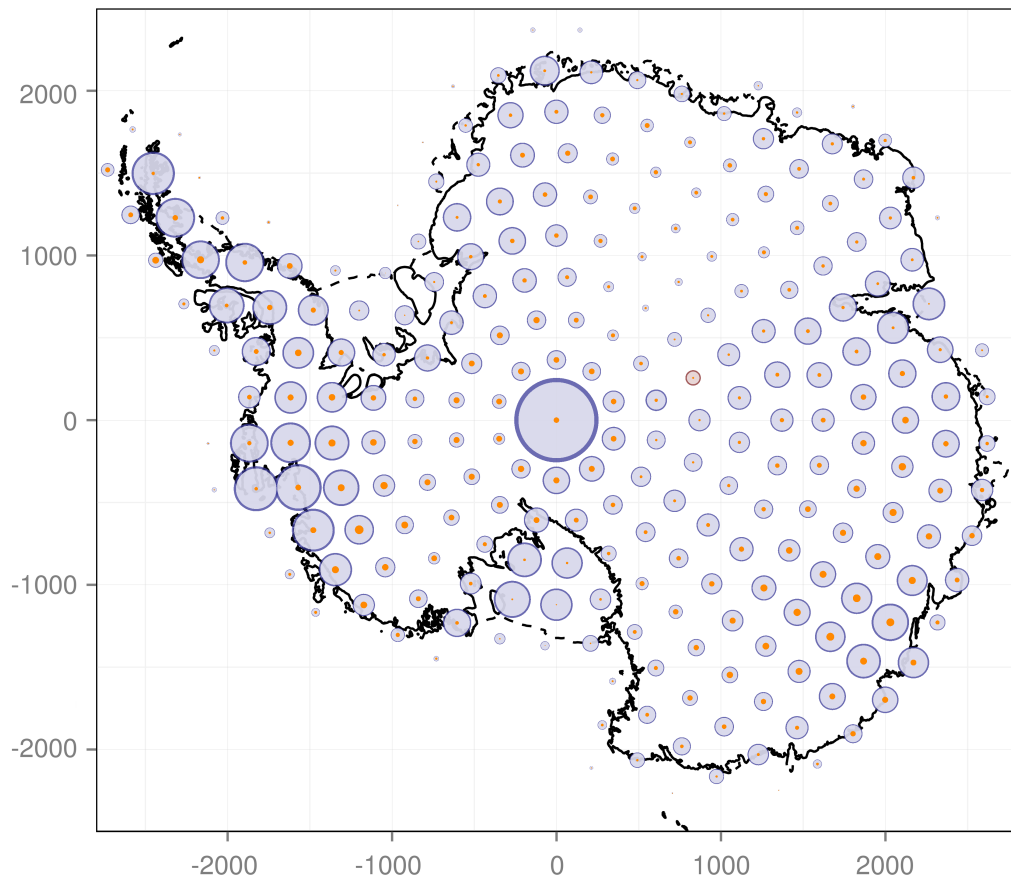


Figure 1: Medal plot for GRACE observations over Antarctica, with distances in kilometres. The solid line is the grounding line (where the ice floats), and the dashed line is the coastline (the limit of ice extent). See section 4 for details of the application and observations, and section 2.4 for the interpretation of the medals. We have used a semi-transparent grey instead of white for the annulus.

ance, because the ice velocity and expected accumulation is so small. Moreover, the rims are all very thin, indicating (in the case of the blue rims) that the initial variance is much larger than the observation error variance. (In fact, in our plotting we expand the rims slightly, where they would otherwise be hardly visible.)

Second, there are clear spatial patterns in the observation error uncertainties. These uncertainties are provided along with the GRACE observations, although we have had to infer a covariance structure. These patterns in the uncertainties are related to physical features that induce variations in the GRACE observations for successive overpasses of the satellites. For example, uncertainty outside the grounding line tends to be small, because the (floating) ice is in hydrostatic equilibrium. Around the grounding line the uncertainty tends to be relatively large because of variations in precipitation and ice velocity. The large medal at the South Pole simply reflects a GRACE observation with a larger spatial footprint, owing to how we have regridded the observations.

Third, the globally updated uncertainties are much smaller than the locally updated uncertainties, as indicated by a thick grey annulus. Therefore much of the reduction in uncertainty at each location is coming from other observations. Some of this will be from other GRACE observations, because GRACE sees height changes in the underlying rock, which has a very long correlation length (i.e. is spatially very stiff). But some of it might also come from the altimetry observations. While we could do separate medal plots to quantify each contribution, in practice we do not have to. Altimetry observations are dominated by surface processes with short correlation lengths. But altimetry satellites cannot overfly the South Pole, and hence there is no altimetry contribution to the South Pole medal. Since the updated variance is about the same at the South Pole as the other medals, we conclude that it is other GRACE observations that dominate the global update shown in each GRACE medal.

These rationalisations of the GRACE medal plot increase our confidence in our statistical modelling and also in our computation. Experienced modellers will appreciate that earlier medal plots of these and the other observations in this application presented apparent anomalies which we were unable to rationalise or verify through testing. We traced these back to modelling or computing choices that we subsequently revisited.

## Acknowledgements

We would like to thank Finn Lindgren and Botond Cseke for several helpful discussions on the computations.

## References

- A. Apte, C.K.R.T. Jones, A.M. Stuart, and J. Voss, 2008. Data assimilation: Mathematical and statistical perspectives. *International Journal of Numerical Methods in Fluids*, **56**, 1033–1046.
- W. Edwards, H. Lindman, and L.J. Savage, 1963. Bayesian statistical inference for psychological research. *Psychological Review*, **70**(3), 193–242.
- A.M. Erisman and W.F. Tinney, 1975. On computing certain elements of the inverse of a sparse matrix. *Communications of the ACM*, **18**(3), 177–179.
- M. Goldstein and D.A. Wooff, 2007. *Bayes Linear Statistics: Theory & Methods*. John Wiley & Sons, Chichester, UK.
- G.H. Golub and C.F. Van Loan, 1996. *Matrix Computations*. Johns Hopkins University Press, Baltimore MD, USA, 3rd revised edition.
- E. Kalnay, 2002. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, Cambridge, UK.
- A. Krause, A. Singh, and C. Guestrin, 2008. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, **9**, 235–284.
- F. Lindgren, H. Rue, and J. Lindström, 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B*, **73**(4), 423–498.
- K.V. Mardia, J.T. Kent, and J.M. Bibby, 1979. *Multivariate Analysis*. Harcourt Brace & Co., London.
- R. Piziak and P.L. Odell, 2007. *Matrix Theory: From Generalized Inverses to Jordan Form*. Chapman & Hall/CRC, Boca Raton FL, USA.
- H. Rue and L. Held, 2005. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton FL, USA.

- H. Rue and S. Martino, 2007. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference*, **137**, 3177–3192.
- L.J. Savage *et al.*, 1962. *The Foundations of Statistical Inference*. Methuen, London.
- J.Q. Smith, 2010. *Bayesian Decision Analysis: Principle and Practice*. Cambridge University Press, Cambridge, UK.
- A. Zammit-Mangion, J.C. Rougier, J. Bamber, and N. Schoen, 2014a. Resolving the Antarctic contribution to sea-level rise: A hierarchical modelling framework. *Environmetrics*, **25**, 245–264.
- A. Zammit-Mangion, J.C. Rougier, N. Schoen, and J. Bamber, 2014b. Multivariate spatio-temporal modelling for assessing Antarctica’s present-day contribution to sea-level rise. Work in progress.